コーポレート・ガバナンス報告書における機械翻訳の検討

2019年 5月 8日

株式会社日本取引所グループ

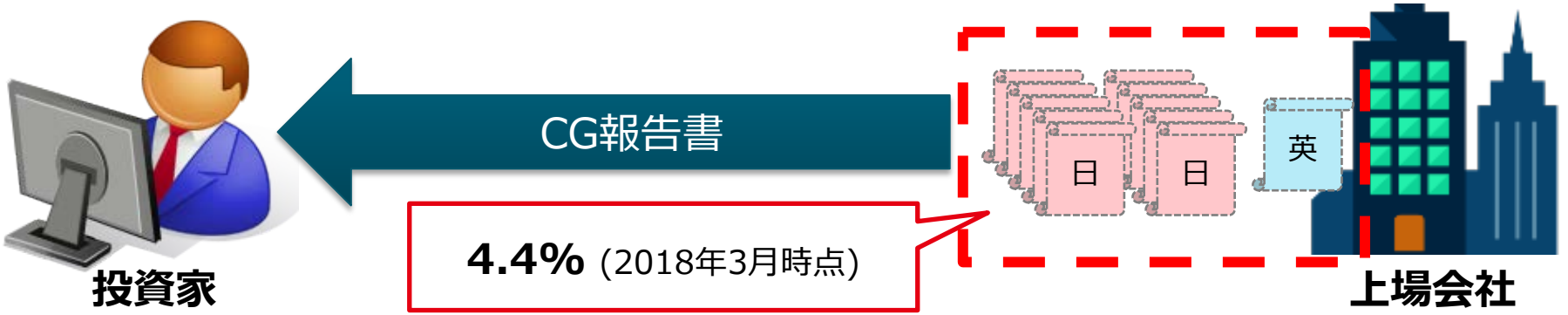
土井 惟成[1], 近藤 真史[2], 山藤 敦史[3]

1. 株式会社日本取引所グループ 総合企画部 フィンテック推進室 部員 (n-doi [at] jpx.co.jp)
2. 株式会社日本取引所グループ 総合企画部 フィンテック推進室 調査役 (m-kondo [at] jpx.co.jp)
3. 株式会社日本取引所グループ 総合企画部 フィンテック推進室 室長 (a-santo [at] jpx.co.jp)

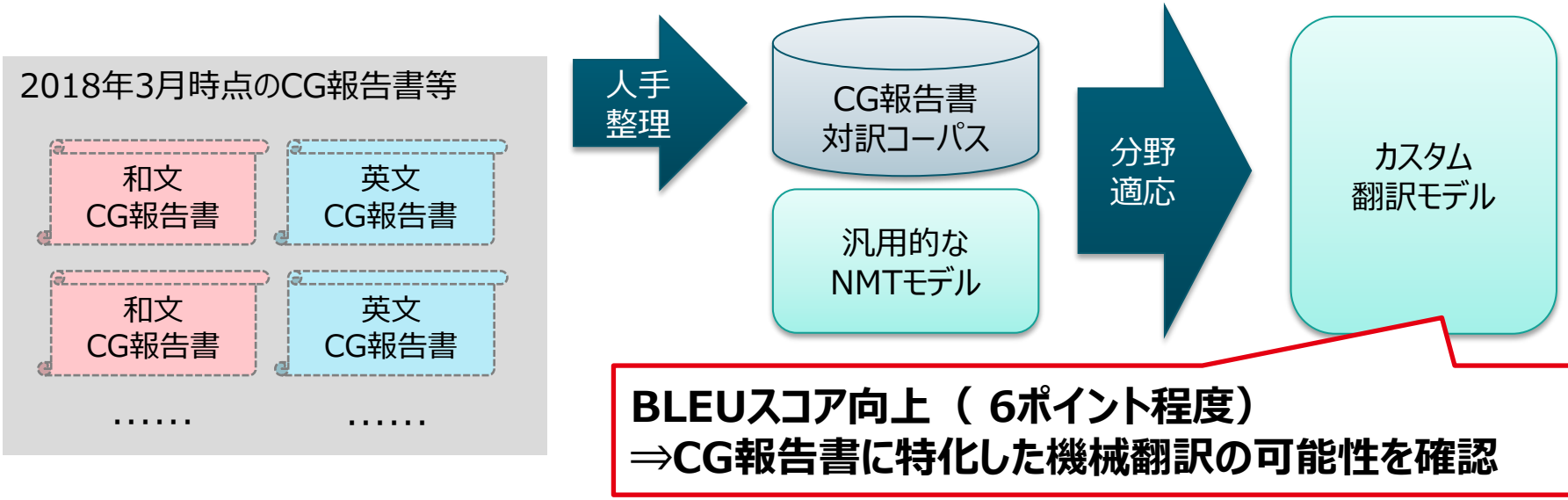
本稿は言語処理学会第25 回年次大会(NLP2019) 発表論文集(2019 年3 月) に掲載の「コーポレート・ガバナンス報告書における機械翻訳の検討」を加筆及び修正したものです。

本研究の概要

■ **コーポレート・ガバナンスに関する報告書（CG 報告書）**の英訳は少ない



■ CG報告書に特化した**カスタム翻訳モデル**の翻訳性能を評価することで、CG報告書における機械翻訳の活用可能性について検討



BLEUスコア向上（ 6ポイント程度）
⇒CG報告書に特化した機械翻訳の可能性を確認

東証の紹介

- 2019年3月現在、約3,600社が上場
- 投資家は東証を通じて上場会社の株式の売買が可能

- 海外投資家の存在感は継続して上昇中
  - 海外投資家による保有比率（時価総額ベース）：**30%超**（2018年3月末時点）
  - 海外投資家が売買代金に占める割合：**60%超**（2018年5月末時点）

投資家

上場会社

売買（投資）

上場

東証

東証の紹介

- 2019年3月現在、約3,600社が上場
- 投資家は東証を通じて上場会社の株式の売買が可能

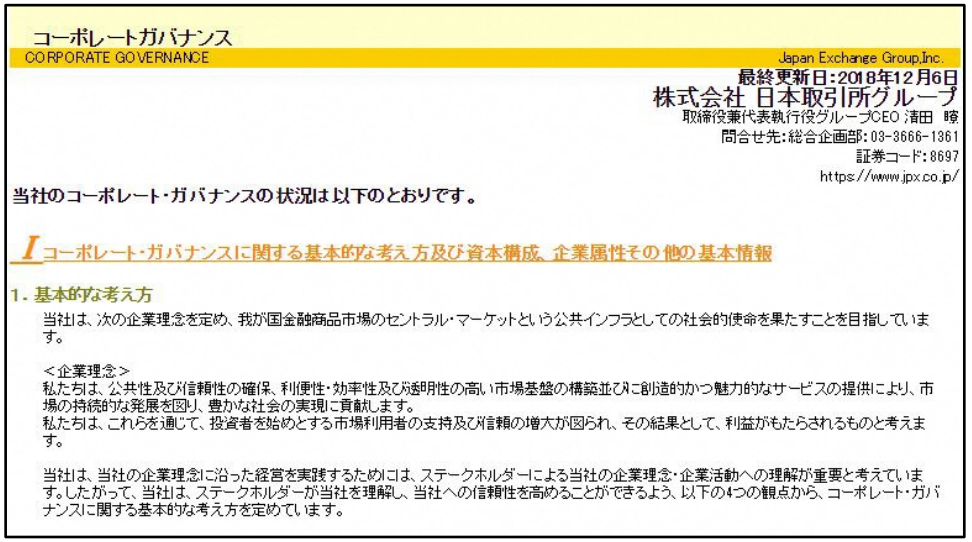
- 上場会社は投資家に対し、**投資判断を行う上で必要な会社情報**を、**公平かつ適時・適切**に開示することが必要 **（適時開示制度）**

英文による会社情報の開示

- 日本の証券市場において、海外投資家の存在感が高まっているものの、**和文と英文の両方の適時開示を行う企業の割合は高くない**
- 投資家にとって重要な情報源の一つである**「コーポレート・ガバナンスに関する報告書」（CG 報告書）**を英文でも開示している上場会社は、全体の**約4.4%**

コーポレート・ガバナンスに関する報告書（CG報告書）

- 上場会社のコーポレート・ガバナンスの状況を投資家に伝えるための書類
- **XBRL**（**XML**を基にして構築されたマークアップ言語）によって記述



コーポレートガバナンス
CORPORATE GOVERNANCE

Japan Exchange Group,Inc.

最終更新日：2018年12月6日
株式会社 日本取引所グループ
取締役兼代表執行役グループCEO 清田 瞭
問合せ先：総合企画部：03-3666-1361
証券コード：8697
https://www.jpx.co.jp/

当社のコーポレート・ガバナンスの状況は以下のとおりです。

Ⅰ コーポレート・ガバナンスに関する基本的な考え方及び資本構成、企業属性その他の基本情報

1．基本的な考え方

当社は、次の企業理念を定め、我が国金融商品市場のセントラル・マーケットという公共インフラとしての社会的使命を果たすことを目指しています。

＜企業理念＞
私たちは、公共性及び信頼性の確保、利便性・効率性及び透明性の高い市場基盤の構築並びに創造的かつ魅力的なサービスの提供により、市場の持続的な発展を図り、豊かな社会の実現に貢献します。
私たちは、これらを通じて、投資者を始めとする市場利用者の支持及び信頼の増大が図られ、その結果として、利益がもたらされるものと考えます。

当社は、当社の企業理念に沿った経営を実践するためには、ステークホルダーによる当社の企業理念・企業活動への理解が重要と考えています。したがって、当社は、ステークホルダーが当社を理解し、当社への信頼性を高めることができるよう、以下の4つの観点から、コーポレート・ガバナンスに関する基本的な考え方を定めています。

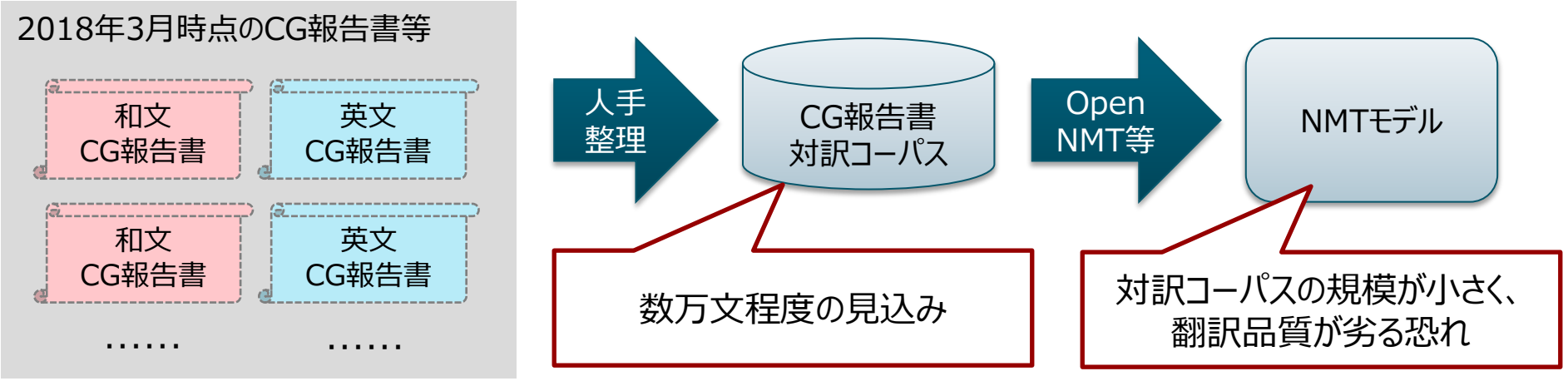
CG報告書の例

```
<tse-t-cg:FinalUpdate contextRef="CG">
  2018-12-06
</tse-t-cg:FinalUpdate>
<tse-t-cg:CompanyName contextRef="CG">
  株式会社 日本取引所グループ
</tse-t-cg:CompanyName>
<tse-t-cg:CompanyNameEn contextRef="CG">
  Japan Exchange Group,Inc.
</tse-t-cg:CompanyNameEn>
<tse-t-cg:RepresentativeSTitleAndName contextRef="CG">
  取締役兼代表執行役グループCEO 清田　瞭
</tse-t-cg:RepresentativeSTitleAndName>
<tse-t-cg:Contact contextRef="CG">
  総合企画部：03-3666-1361
</tse-t-cg:Contact>
<tse-t-cg:URL contextRef="CG">
  https://www.jpx.co.jp/
</tse-t-cg:URL>
<tse-t-cg:BasicPolicy contextRef="CG">
当社は、次の企業理念を定め、我が国金融商品市場のセントラル・マーケット
という公共インフラとしての社会的使命を果たすことを目指しています。
```

XBRLの例

CG報告書の機械翻訳における問題点

- CG報告書特有の専門用語・固有名詞・文体が存在し、汎用的なNMTサービスでは正確な翻訳が困難（詳細後述）
- CG報告書のみでは、対訳コーパスを作成するには規模が小さくなる見込み



本研究の手法

- **カスタム翻訳モデル(＊)**に対応した機械翻訳サービスに、既存のCG報告書等から作成した対訳コーパスを用いて、CG報告書に適したNMTモデルを作成

(*) 汎用的なNMTモデルに対して**分野適応（Domain Adaptation）**を施したNMTモデル

本研究の流れ

Step1：予備実験

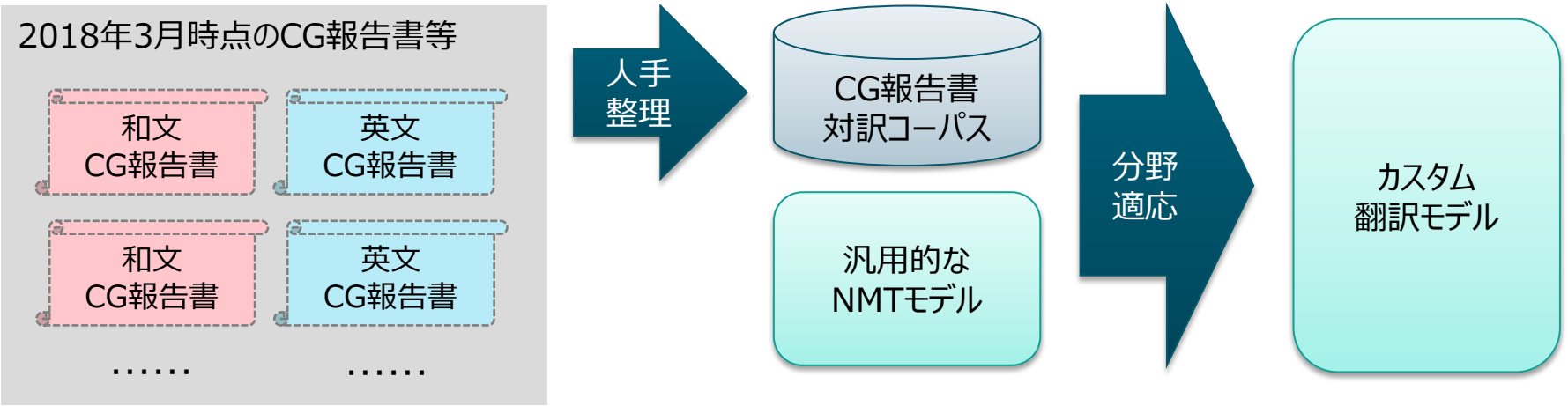
- CG報告書の機械翻訳における問題点を整理
- 誤訳が生じやすい和文の傾向などの調査

Step2：対訳コーパス作成

- 既存のCG報告書を用いて対訳コーパスを作成
- 重複文の除去をはじめとする前処理を実施

Step3：実験・評価

- 汎用的なNMTモデルに対して分野適応を施したカスタム翻訳モデルを作成
- その翻訳品質を評価することで、CG 報告書における機械翻訳の可能性を検証



誤訳しやすい和文の傾向

1. 専門用語・固有名詞	•専門用語及び固有名詞が多数出現 •専門用語によっては代表的な英訳が一意に定まっていない
2. 訳揺れしやすい代名詞	•「当社」や「同氏」といった代名詞が多数出現 •これらの英訳には複数の候補が存在
3. 一文中の並列句	•項目の細分化や項目の列挙を行うに当たって、一文中に並列句を挿入することがある
4. 名詞の非限定列挙	•「甲、乙、丙等」における「等」 •正確な英訳には前後の文脈やその単語の前提知識が必要
5. 見出し符号 (見出し番号)	•(1) (2) (3) や a) b) c) •(イ)には『アイウ(2番目)』か『イロハ(1番目)』の2通りが存在
6. 造語・詩的な表現	•日本語特有のニュアンスを踏まえた造語や詩的な表現

傾向３：一文中の並列句

原文

強化すべき点としましては、当社事業がグローバルに拡大する中において、ガバナンス機能の更なる充実に向けた取組みが重要との認識に立ち、

- 監督機能をより強化するため、取締役会で意思決定した重要な事項に対する継続的なモニタリングをより充実させること
- リスクマネジメント視点での議論を更に強化すること

などについて、取り組むこととしました。

一行ずつ機械翻訳

（最終行）
I decided to work on what.

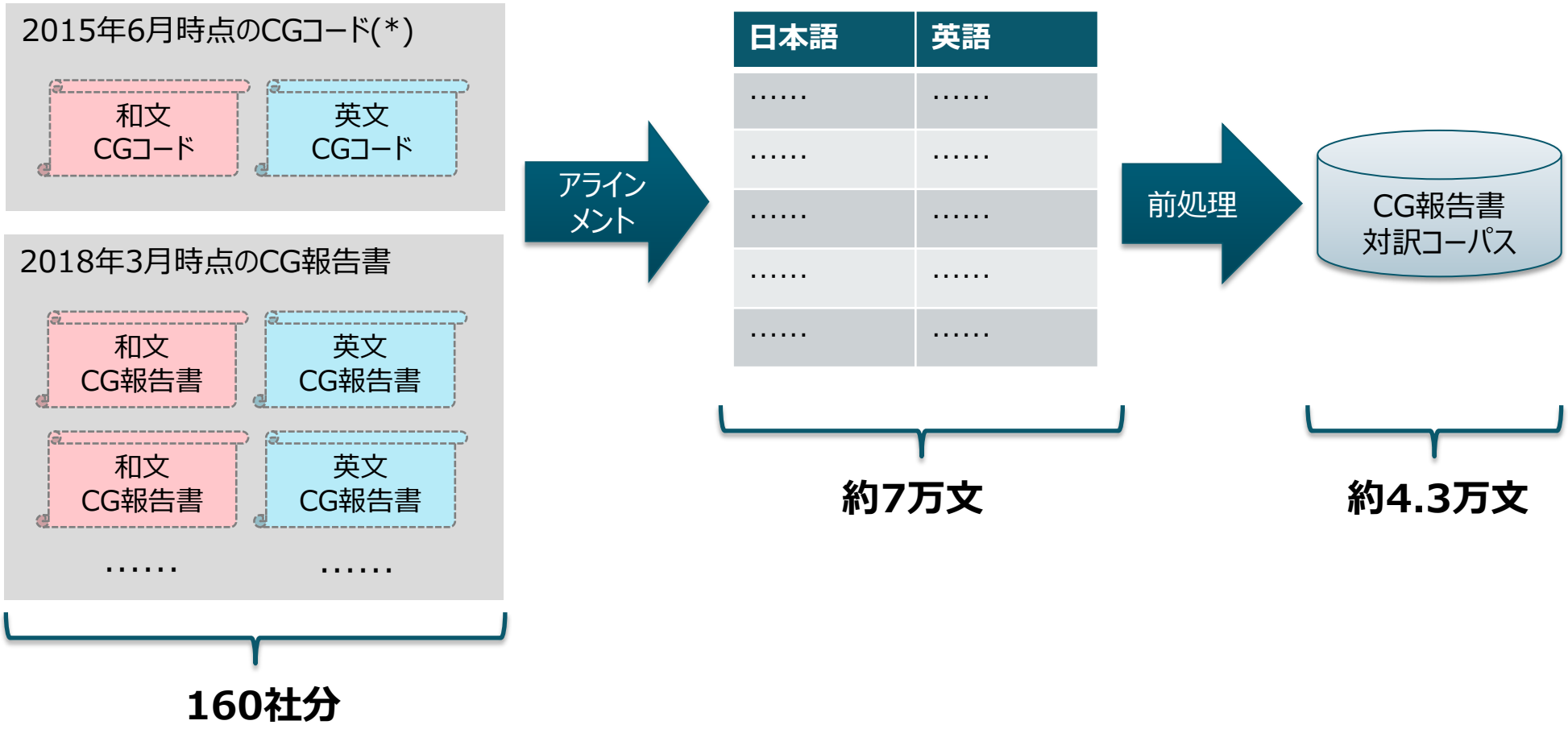
参照訳

Recognizing that efforts for further enhancement of the governance function is vital to Santen in the midst of the global expansion of its business, Santen has decided to exert further efforts with respect to items that need to be strengthened including those **listed below:**

- With the aim of strengthening the monitoring function, further enhancing the continuous monitoring of material matters that are decided at meetings of the Board of Directors; and
- Further strengthening discussions from the viewpoint of risk management.

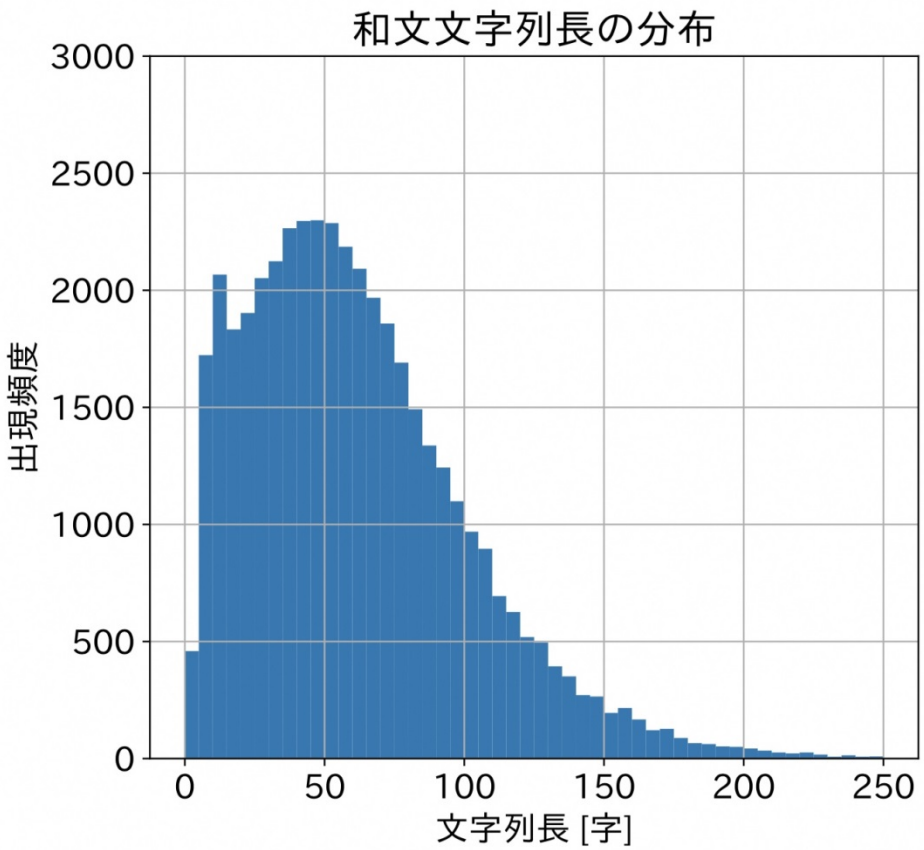
CG報告書対訳コーパスの作成

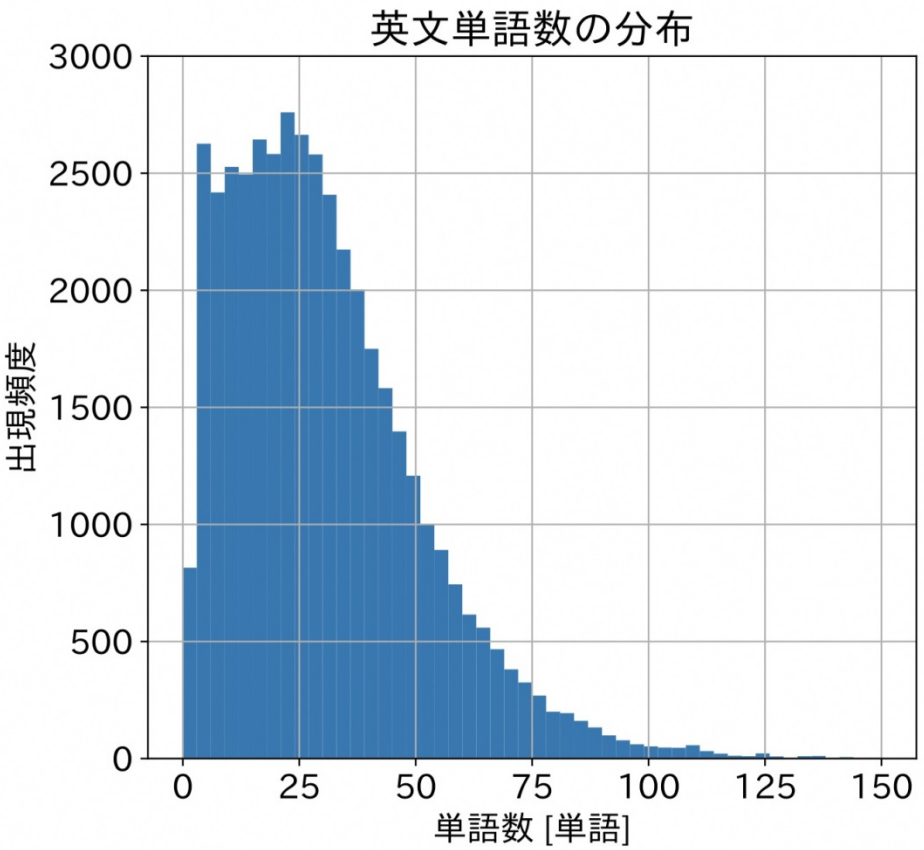
■ CG報告書対訳コーパス作成の流れ



(*) コーポレートガバナンス・コード。実効的なコーポレート・ガバナンスの実現に資する主要な原則を取りまとめた規則。

和文文字列長の分布と英文単語数の分布





実験設定

■ 実験環境
- AutoML Translation（Google Cloud Platform™）
- 他サービスA（カスタム翻訳モデルの作成を提供しているクラウド型の機械翻訳サービス）

■ 実験データ
- CG報告書対訳コーパス；全量版
- CG報告書対訳コーパス；長文除去版
  - ✓ 和文文字列長100文字以上 or 英文単語数50単語以上 の対訳を削除
  - ✓ サービスAでは英文単語数50単語以上の対訳等が自動的に削除されるため、全量版のみ利用

サービス	データ	学習用	開発用	評価用	合計
AutoML Translation	全量版	34,527	4,316	4,315	**43,158**
	長文除去版	27,641	3,455	3,455	**34,551**
		80% : 10% : 10%			
サービスA	全量版	38,446	2,129	2,033	**42,704**
		90% : 5% : 5%			

実験結果

■ 定量評価方法

- BLEUによる評価
- 3回平均の値を算出

■ 評価結果

サービス	データ	カスタム版	ベースライン	上昇量
AutoML Translation	全量版	25.57	19.63	**+5.94**
	長文除去版	26.02	19.78	**+6.24**
サービスA	全量版	28.27	21.62	**+6.66**

得られた示唆

- いずれの場合でもBLEUスコアの有意な上昇を確認
- カスタム翻訳モデルによって、全体としては翻訳品質が向上したことを推察

訳出結果の分析

■ 専門用語・固有名詞の翻訳は改善

1. 専門用語・固有名詞	•専門用語及び固有名詞が多数出現 •専門用語によっては代表的な英訳が一意に定まっていない

■ **前後の文脈や各単語の前提知識が必要となる文**については、一文ごとに機械翻訳を行うNMTモデルにおける分野適応では解決が難しい

2. 訳揺れしやすい代名詞	•「当社」や「同氏」といった代名詞が多数出現 •これらの英訳には複数の候補が存在
4. 名詞の非限定列挙	•「甲、乙、丙等」における「等」 •正確な英訳には前後の文脈やその単語の前提知識が必要
5. 見出し符号 (見出し番号)	•(1) (2) (3) や a) b) c) •(イ)には『アイウ(2番目)』か『イロハ(1番目)』の2通りが存在
6. 造語・詩的な表現	•日本語特有のニュアンスを踏まえた造語や詩的な表現

改善例

入力文	株式会社ＩＣＪの提供する**議決権電子行使プラットフォーム**に参加しております。
参照訳	The Company is participating in **the platform for the electronic exercise of voting rights** operated by ICJ, Inc.
ベースライン	We participate in **the platform for exercising voting rights electronics** provided by ICJ Co., Ltd.
カスタム版	The Company participates in **the electronic voting platform** provided by ICJ, Inc.

入力文	２０１６年度において、当社の会計監査人は**新日本有限責任監査法人**であります。
参照訳	The Accounting Auditor of the Company for the year ended March 31, 2017 is **Ernst & Young ShinNihon LLC**.
ベースライン	In fiscal 2016, our accounting auditor is **New Japan limited liability audit corporation**.
カスタム版	In fiscal 2016, the Company’s Accounting Auditor was **Ernst & Young ShinNihon LLC**.

入力文	当社グループは、企業価値の最大化を目指し、コーポレート・ガバナンスの徹底を**最重要課題**の一つと位置付け様々な施策を講じています。
参照訳	The Group is aiming to maximize corporate value, and has been implementing various measures as rigorous corporate governance is **our highest priority**.
ベースライン	The Group aims to maximize corporate value and positions thorough corporate governance as one of **the most important tasks** and takes various measures.
カスタム版	The Group considers thoroughness of corporate governance as one of **the most important issues** with the aim of maximizing corporate value, and takes various measures.

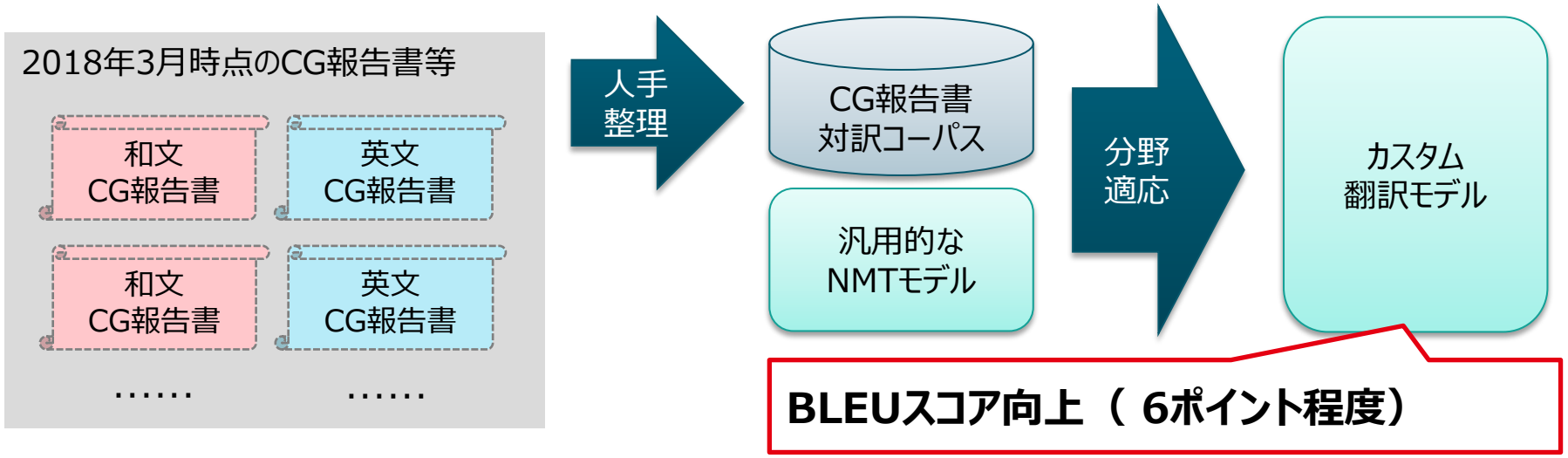
改善しなかった例

入力文	さらに**同氏**は取締役会議長として、当社グループの経営の基本方針等について、取締役会としての決議に向け議案審議を主導いたしました。
参照訳	Furthermore, as the Chairman of the Board of Directors, **she** led the Board of Directors to make decisions on proposals, including a proposal for basic management policy of the Group.
ベースライン	In addition, as Chairman of the Board of Directors, **he** led the deliberation on the agenda for resolutions as the Board of Directors regarding the Group's basic management policies and others.
カスタム版	In addition, as Chairman of the Board of Directors, **he** led the deliberation of the Board of Directors on the basic policy on the management of the Group.

入力文	外部識者による講演会の開催、社内WEB サイトでの情報発信、座談会実施**等による啓発活動**
参照訳	**Holding awareness campaigns through** round-table discussions, publishing of information on the Company's internal website, and hosting lectures by visiting experts.
ベースライン	Held lectures by outside experts, disseminate information on internal website and **raise awareness through** implementation of round-table discussion etc.
カスタム版	Held lectures by outside experts, disseminate information on the internal website, and **conduct awareness-raising activities such as** holding round-table talks.

まとめ

- 既存のCG 報告書等を用いて対訳コーパスを構築し、それを用いたカスタム翻訳モデルを評価することで、CG 報告書における機械翻訳の活用可能性を検証
- 実験により、**専門用語や固有名詞**をはじめとする英訳の改善によって、BLEU スコアの有意な上昇を確認
- 本研究で構築した対訳コーパスは比較的少量であることから、**この拡充と品質の向上によって、翻訳品質のさらなる向上を期待**



- ただし、**前後の文脈や前提知識等を踏まえた翻訳が必要な文**については、これらを考慮する機械翻訳モデルや、前処理や後処理等が必要であると推察

今後の課題

■ 機械翻訳の課題

- 文脈を考慮する機械翻訳モデルの検討
- 見出し符号(見出し番号) 等における前処理及び後処理の検討
- 記載内容に応じた機械翻訳モデルの検討
  - ✓ 例: XML 形式のタグ情報から，短文・長文・固有名詞などの属性を推定し，ルールベース機械翻訳モデルやNMT モデルをはじめとする機械翻訳モデルの使い分け

■ 外部の課題

- 証券・IR 分野における対訳コーパスの品質向上及び拡充
- 証券・IR 分野における用語集及びスタイルガイド等の検討
- 特定分野の対訳コーパスの構築における，対応付け(アライメント) ・前処理・データクレンジングのガイドラインの検討
- NMT モデルに適した和文の書き方等の体系的な調査

参考文献

1. 土井惟成, 近藤真史, 山藤敦史. コーポレート・ガバナンス報告書における機械翻訳の検討. 言語処理学会第25 回年次大会(NLP2019), pp. 926–929, 3 2019.
2. 株式会社東京証券取引所, 株式会社名古屋証券取引所, 証券会員制法人福岡証券取引所, 証券会員制法人札幌証券取引所. ２０１７年度株式分布状況調査結果の概要.
3. 株式会社東京証券取引所情報サービス部. 投資部門別売買状況. https://www.jpx.co.jp/markets/statistics-equities/investor-type/00-02.html, 2017. (参照2019-01-15).
4. 投資家フォーラム. 投資家フォーラム-第１・２回会合-報告書. https://investorforum.jp/. (参照2019-01-15).
5. 株式会社東京証券取引所. 東証上場会社コーポレート・ガバナンス白書, 2017.
6. Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. InProceedings of the 27th International Conference on Neural Information Processing Systems - Volume2, NIPS'14, pp. 3104–3112, Cambridge, MA, USA, 2014. MIT Press.
7. Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In Proceedingsof the First Workshop on Neural Machine Translation, pp. 28–39. Association for ComputationalLinguistics, 2017.
8. 株式会社みらい翻訳. 職場英語力をTOEIC 900 点相当に引き上げる日英双方向機械翻訳サービスをリリース. https://miraitranslate.com/uploads/2017/12/befdfe2e9eca64235a2042cd9f50a3db.pdf, 12 2017. (参照2019-01-15).
9. Translation API - Dynamic Translation | Translation API. https://cloud.google.com/translate/. (参照2019-01-15).
10. Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. Bleu: a method for automatic evaluationof machine translation. pp. 311–318, 2002.
11. 藤田篤, 山田優, 影浦峡. 産業翻訳に役立つ自然言語処理技術についての議論の足場. 言語処理学会第25 回年次大会(NLP2019), pp. 914–917, 3 2019.
12. AutoML Translation | Google Cloud. https://cloud.google.com/translate/automl/docs/.(参照2019-01-15).
13. Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey,Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson,Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, KeithStevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick,Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translationsystem: Bridging the gap between human and machine translation. CoRR, Vol. abs/1609.08144,, 2016.